Classification Algorithm in Machine Learning

Classification Algorithm in Machine Learning

As we know, the Supervised Machine Learning algorithm can be broadly classified into Regression and Classification Algorithms. In Regression algorithms, we have predicted the output for continuous values, but to predict the categorical values, we need Classification algorithms.

What is the Classification Algorithm?

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.

In classification algorithm, a discrete output function(y) is mapped to input variable(x).

$y=f(x)$, where y = categorical output

```
y=f(x), where y =
categorical output
```

The best example of an ML classification algorithm is **Email Spam Detector**.

The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.

Classification algorithms can be better understood using the below diagram. In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.
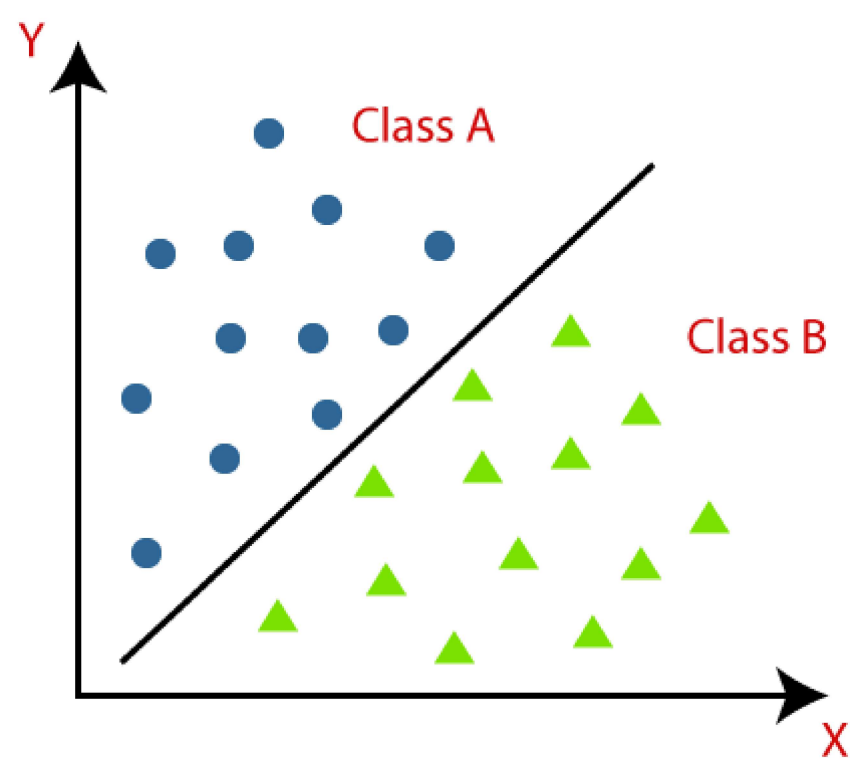
The algorithm which implements the classification on a dataset is known as a classifier. There are two types of Classifications:

Binary Classifier: If the classification problem has only two possible outcomes, then it is called as Binary Classifier.
Examples: YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.
Multi-class Classifier: If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.
Example: Classifications of types of crops, Classification of types of music.

Learners in Classification Problems:

In the classification problems, there are two types of learners:

Lazy Learners: Lazy Learner firstly stores the training dataset and wait until it receives the test dataset. In Lazy learner case, classification is done on the basis of the most related data stored in the training dataset. It takes less time in training but more time for predictions.
Example: K-NN algorithm, Case-based reasoning

Eager Learners: Eager Learners develop a classification model based on a training dataset before receiving a test dataset. Opposite to Lazy learners, Eager Learner takes more time in learning, and less time in prediction.
Example: Decision Trees, Naïve Bayes, ANN.

Types of ML Classification Algorithms:

Classification Algorithms can be further divided into the Mainly two category:

Linear Models
Logistic Regression
Support Vector Machines

Non-linear Models
K-Nearest Neighbours
Kernel SVM
Naïve Bayes
Decision Tree Classification
Random Forest Classification

Note: We will learn the above algorithms in later chapters.

Evaluating a Classification model:

Once our model is completed, it is necessary to evaluate its performance; either it is a Classification or Regression model. So for evaluating a Classification model, we have the following ways:

1. Log Loss or Cross-Entropy Loss:

It is used for evaluating the performance of a classifier, whose output is a probability value between the 0 and 1.

For a good binary Classification model, the value of log loss should be near to 0.

The value of log loss increases if the predicted value deviates from the actual value.

The lower log loss represents the higher accuracy of the model.

For Binary classification, cross-entropy can be calculated as:

$?(y\log(p)+(1?y)\log(1?p))$

```
    -(ylog(p)+(1-
y)log(1-p))
```

Where y= Actual output, p= predicted output.

2. Confusion Matrix:

The confusion matrix provides us a matrix/table as output and describes the performance of the model.

It is also known as the error matrix.

The matrix consists of predictions result in a summarized form, which has a total number of correct predictions and incorrect predictions. The matrix looks like as below table:

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

$$\text{Accuracy} = \frac{TP+TN}{\text{Total Population}}$$

3. AUC-ROC curve:

ROC curve stands for **Receiver Operating Characteristics Curve** and AUC stands for **Area Under the Curve**.

It is a graph that shows the performance of the classification model at different thresholds.

To visualize the performance of the multi-class classification model, we use the AUC-ROC Curve.

The ROC curve is plotted with TPR and FPR, where TPR (True Positive Rate) on Y-axis and FPR(False Positive Rate) on X-axis.

Use cases of Classification Algorithms

Classification algorithms can be used in different places. Below are some popular use cases of Classification Algorithms:

Email Spam Detection

Speech Recognition

Identifications of Cancer tumor cells.

Drugs Classification

Biometric Identification, etc.

Next Topic[Logistic Regression](#)